In'nami *et al. Language Testing in Asia* (2016) 6:3
DOI 10.1186/s40468-016-0025-9

Language Testing in Asia
a SpringerOpen Journal

RESEARCH **Open Access**

Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test

Yo In'nami[1*], Rie Koizumi[2] and Keita Nakamura[3]

* Correspondence:
innami@tamacc.chuo-u.ac.jp
[1]Chuo University, Tokyo, Japan
Full list of author information is available at the end of the article

Abstract

Background: This study examined the factor structure of the Test of English for Academic Purposes (TEAP®) test—a recently developed academic English test measuring four skills among Japanese university applicants—and compared the structure to that of the Test of English as a Foreign Language Internet-based test (TOEFL iBT®), to investigate the extent towhich the TEAP test is related to the TOEFL iBT test.

Methods: Using confirmatory item-level factor analysis and scores on both tests, obtained from 100 students, we tested four models (unitary, correlated, receptive-productive, andhigher-order) for the TEAP test.

Results: We found that the higher-order model fit the data best. This suggests that the TEAP measures the four skills of reading, listening, writing, and speaking well and that they could be conceptualized as reflecting a single academic proficiency. This supports the appropriateness of the constructs, as defined and operationalized in the specifications of the TEAP test. Further, we found a close relationship between the TEAP and the TOEFL iBT tests (1.005). This suggests that both tests measure a very similar construct. This provides positive evidence of the concurrent validity of the TEAP test, as an indicator of academic English skills. These results were also supported by a follow-up analysis using item-parceled data.

Conclusions: The close relationship between the TEAP and TOEFL iBT tests suggests that the TEAP test measures the intended construct of the four skills in academic settings very well.

Keywords: Factor structure, TEAP, TOEFL iBT, Validation, Concurrent validity

Background

English is the lingua franca in the public, academic, and corporate sectors in many parts of the world, so much so that the importance of English skills for students cannot be overstated. Japan is no exception to this global trend, and a variety of standardized tests for measuring learners' English skills are available in the country: the Test of English as a Foreign Language Internet-based Test (TOEFL iBT®; Educational Testing Service 2015a), TOEFL Junior® Comprehensive (Educational Testing Service 2015b), Test of English for International Communication (TOEIC; Educational Testing Service 2015d), International English Language Testing System (IELTS; British Council, IDP, IELTS Australia, & Cambridge English Language Assessment, n.d.), Cambridge English exams (University of Cambridge Local Examinations Syndicate 2015), and the Pearson Test of English Academic (Pearson 2014), all of which are known worldwide; and the Test in Practical English Proficiency (EIKEN; Eiken Foundation of Japan, n.d.a.), Global Test of English Communication (GTEC) for STUDENTS (Benesse 2000–2015b), GTEC Computer Based Testing (CBT; Benesse 2000–2015a), and the Computerized Assessment System for English Communication (CASEC; Japan Institute for Educational Measurement, n.d.), all of which target the domestic market. The use of these tests to screen applicants for Japanese universities has been currently discussed in Japan, as improvement in university entrance examinations would strengthen the positive relationships between language teaching, learning, and assessment.

In this article, we focused on Japanese university entrance examinations and report on the Test of English for Academic Purposes (TEAP®) test—a recently developed English test for Japanese university applicants, designed to measure four skills, namely, reading, listening, writing, and speaking in an academic context. As part of its validation study, we report on the factor structure of the TEAP test in relation to that of the TOEFL iBT test—a more established instrument of the four skills in an academic setting—to examine the extent to which the TEAP test measures a construct similar to the one measured by the TOEFL iBT test. We decided to compare the TEAP test with the TOEFL iBT test because the TOEFL iBT measures the four skills of reading, listening, writing, and speaking in an academic setting and its validity evidence has been widely investigated (e.g., Chapelle et al. 2008; see also TOEFL research reports published by Educational Testing Service 2015c).

Test of English for Academic Purposes

The Test of English for Academic Purposes (TEAP: pronounced /tíːp/) was recently developed by the Eiken Foundation of Japan and Sophia University—a private university in Tokyo known for its English language program, in collaboration with the Centre for Research in English Language Learning and Assessment (CRELLA) of the University of Bedfordshire in the United Kingdom. It was constructed with specific use in mind—English examination for university entrance in Japan. The test aims to measure a construct—academic English proficiency required for learning and researching at universities. The TEAP test measures four skills (summarized in Table 1). The Reading Section and Listening Section are designed to measure mainly the understanding of short and long passages with visual information. The Speaking Section is face-to-face, one-on-one interviews, and includes both monologue and dialogue tasks on various issues. The Writing Section includes both summary and integrated tasks. The TEAP test is conducted thrice a year (in July, September, and December) at 11 major cities in Japan, as of 2015.

The TEAP test covers the A2 through B2 levels of the Common European Framework of Reference for Languages (CEFR), the lowest among which is the A2 level. Learners at this level can understand simple information and communicate in familiar situations, for example, participating in a straightforward conversation on everyday topics. The B1 level refers to learners with the ability to deal with routine matters at school, work, or when traveling, etc., for example, understanding the main point of a simple text or asking how to borrow a book at the library. Finally, the B2 level learners

Table 1 Four Sections of the TEAP Test

Part	Skill	*k*
	Reading (70 min; multiple choice)	60
1	Vocabulary and word usage	20
2A	Understanding graphs and charts	5
2B	Understanding notices and e-mails	5
2C	Understanding short texts	10
3A	Understanding long texts	8
3B	Understanding long texts (including graphs and charts)	12
	Listening (50 min; multiple choice)	50
1A	Understanding short conversations	10
1B	Understanding short talks, news, and lectures	10
1C	Understanding short texts, including graphs and charts	5
2A	Understanding long conversations, including three-way interactions	9
3A	Understanding long passages such as lectures, including graphs and charts	16
	Speaking (10 min; face-to-face, one-on-one interview)	4
1	Describing oneself	1
2	Role-playing: Interviewing an interviewer (tests the ability to lead conversations)	1
3	Making a speech on current issues	1
4	Responding to questions on current issues	1
	Writing (70 min; 2 tasks)	2
1	Summarizing an expository text or critique in about 70 words	1
2	Reading multiple texts, graphs, and charts, summarizing main points, and writing a 200-word opinion essay	1

Note: *k* = number of items/tasks. This table was sourced from the Eiken Foundation of Japan (n.d.b.)

are capable of understanding more complex texts and expressing themselves on more technical topics, for example, understanding the main point of a newspaper article or explaining opinions supported by reasons. Overall, these descriptions suggest that the TEAP targets learners of the beginning and intermediate levels. For further information on the CEFR, please see Council of Europe (2014).

While practice items are available online (Eiken Foundation of Japan, n.d.b.), actual items are IRT-equated for the Reading Section and Listening Section and are, thus, not released to the public. The papers are collected upon completion of the test and cannot be taken home by the examinees. Scores, bands, and TEAP Can-do statements for all of the four sections are provided to examinees.

The TEAP test was introduced in 2014. As of January 2016, 21 Japanese universities have decided to include the test as one of the requirements for applicants. This is remarkable, given that Sophia University—a co-developer of the TEAP test—was the only university to have included the test in its entrance examination system in 2014. Sophia announced that, starting from 2014, it would accept TEAP scores, in addition to those of Eiken, IELTS, TOEIC, TOEFL iBT, and the United Nations Association's Test of English, for its high school recommendation-based exam, third-year transfer exam, and exam for university graduates wishing to be admitted to the Department of Nursing at the Faculty of Human Sciences (Sophia University, n.d.b.). Sophia also announced that it would add a TEAP-based entrance examination system in 2015: Applicants must take the TEAP test beforehand and submit the score to Sophia. Each department in the university sets a minimum TEAP score for a student to qualify for entrance exams. Only students whose score exceeds the minimum score can apply for the university's entrance exams, and scores above the cutoff are treated equally, for which reason high scores exceeding the cutoff are of no advantage (Sophia University, n.d.a.).

One may wonder where the need to develop the TEAP emerged, since currently, both domestic and international English tests are available in Japan. One of the currently major domestically available tests used for selecting applicants for universities is the National Center Test for University Admissions (often called "the Center Test"). It is also possible to take international tests such as the TOEFL iBT, IELTS, Pearson Test of English Academic, Cambridge English exams, and the TOEFL Junior® Comprehensive, which may serve a similar purpose.

As for the Center Test, this test score is currently a major factor in the selection of candidates in national, public, and many private universities. More than half a million students take this test annually (559,132 in 2015; National Center for University Entrance Examinations 2015). The test is administered nationwide only once a year (in mid-January), with local universities as test sites. In some universities, the Center Test score is the only basis for admission; other universities also require candidates to take institution-specific exams, such as an additional English test, essay, and/or interview, but writing and speaking components are rarely administered. The English section of the Center Test aims to measure a wide range of skills—from knowledge of pronunciation, accent, and grammar to reading and listening comprehension—but, one of the major problems is that the Center Test does not directly assess speaking and writing skills. It has been argued that skill imbalance in university entrance examinations tends to have negative washback effects on English learning and teaching at secondary levels (e.g., Shimomura 2014). The four-skill TEAP test was developed to fill this void and is one of the candidates to be used in the replacement of the Center Test. The Central Council for Education (MEXT 2014a)—an advisory board for Japan's Ministry of Education—submitted a report to the Ministry advising that the Center Test be replaced with a four-skill exam in 2020, as part of radical reforms of secondary and higher education and university entrance examinations.

Regarding international tests, the TOEFL iBT, IELTS, and Pearson Test of English Academic could generally be too difficult for most Japanese university applicants. Although empirical evidence for their performance on these measures has been scarce, one of the most relevant studies is MEXT (2014b), which reports the test results of 70,000 randomly sampled third-year Japanese national or public high school students (Grade 12). Four-skill exams aligned with the CEFR were administered, and the results showed that all skills were at the A1 level. With this in mind, the A1 and A2 levels are too low in proficiency to correspond to any scores for reading and listening, to 11 scores for writing (maximum 30 points), or to 13 or 19 scores for speaking (maximum 30 points) in the TOEFL iBT test (see Table 18 in Tannenbaum and Wylie 2008). This suggests that average Japanese high school students would score 24 (0 + 0 + 11 + 13w) to 30 (0 + 0 + 11 + 19) in the TOEFL iBT. As the TOEFL iBT consists of four sections, each of which measures one of the four skills (4 sections * 30 points = maximum 120 points), 24 to 30 of the 120 points are low, suggesting that the TOEFL iBT might be difficult for Japanese university applicants. Second, the IELTS tests could work well for advanced Japanese learners of English; further, other tests such as the TOEFL Junior® Comprehensive and Cambridge English exams could also serve lower-level applicants.

However, they are not necessarily designed with the Japanese high school curriculum guidelines in mind. This is certainly not a fault in these exams, as they are internationally used proficiency exams and not based on any specific country's school curriculum. Using these tests could still have a positive washback on Japanese learners (particularly, on those planning to study abroad), but may also have a negative washback effect, since there is a gap between what they are taught and what they are tested on. We do not mean to criticize these tests. Rather, we intend to underscore that the purposes of these tests are not to screen Japanese university applicants, and that they are not designed specifically for Japanese university applicants. As will be reviewed below, the design of the TEAP test was informed by the results of language function surveys based on Japanese school curriculum guidelines that were taken by Japanese high school teachers and university instructors. The TEAP test is aimed at Japanese university applicants. It is intended to promote more positive relationships among language teaching, learning, and assessment in the Japanese educational context than the other tests currently in use.

Validation studies on the TEAP test

The TEAP test also includes writing and speaking, as well as reading and listening, in the hope that targeting a wider range of relevant and representative constructs in an academic setting will have a positive impact on learning at the secondary level. As the TEAP test can be used in high-stakes university admissions contexts, validation studies have been conducted and reported on the website of the Eiken Foundation of Japan (Eiken Foundation of Japan, n.d.c.), including testing for inter- and intra-rater reliability and the contextual variables on the Reading and Listening Sections (Taylor 2014), and the appropriateness of tasks and rating scales in the Writing (Weir 2014) and Speaking Sections (Nakatsuhara 2014; Nakatsuhara et al. 2014). Stakeholders' perceptions of university entrance examinations and the washback expected from the introduction of the TEAP test have also been investigated (Green 2014; Nakamura 2014). For example, Nakatsuhara (2014) describes the process of developing a draft of the specifications of the TEAP speaking, based on the results of language function surveys that were based on Japanese high school curriculum guidelines and were answered by 167 Japanese high school teachers and 24 Sophia University instructors. A one-day focus group meeting was also held to discuss key issues on the test specifications (e.g., test purposes, target language use, and task types). Also described is the process of developing a rating scale, by referring to the CEFR descriptors and major rating scales such as the Cambridge ESOL Common Scale for Speaking, and those developed for speaking tests aimed at Japanese EFL learners (the Standard Speaking Test [ALC Press 2015], and the Kanda English Proficiency Test—a group oral test that was developed at Kanda University of International Studies in Japan and assesses performance with five analytic scales of pronunciation, fluency, grammar, vocabulary, and communicative strategies [see e.g., Ockey 2009]). The results of her validation study were positive, overall: For example, the transcribed video-recorded performance of 23 students recruited at Sophia University revealed the language functions that the TEAP project team intended to elicit in four tasks. This provided positive validity evidence that the speaking tasks of the TEAP test functioned as intended and that the definition and operationalization of the speaking construct in the test specification seemed to be appropriate.

As validation is an iterative process to accumulate validity evidence in order to make a convincing validity argument (e.g., Chapelle et al. 2008), more validation studies need to be conducted. This is particularly true for the TEAP test, for two additional reasons. First, it should be noted that all validation studies on the TEAP test we reviewed were conducted by researchers at the CRELLA or Eiken Foundation of Japan, namely, by collaborators on or developers of the TEAP test, respectively, and may therefore lack external validity. As the TEAP test was only introduced in 2014, it is understandable but not excusable that all studies were conducted by those associated with its design. Unfortunately, we have not been able to locate validation studies on the TEAP test that were conducted by external researchers. The TEAP developer is aware of this, which is why they commissioned the first and second authors of the current manuscript—external researchers—to conduct this study.

Second, some areas of interest have not been researched yet. One such area is the factor structure of the TEAP test. This gap needs to be filled, because studies on factor structure show whether there are empirically supported relationships between the intended interpretation of scores and the constructs being measured (e.g., Bae and Bachman 1998; Bollen 1989; In'nami and Koizumi 2012; Messick 1996; Sawaki et al. 2009). For example, In'nami and Koizumi (2012) examined the factor structure of the revised TOEIC test and reported that the division of listening and reading skills into different factors supported the reporting of separate scores for each skill (the standardized factor loadings from the listening factor to the listening items were .66 to .83, and those from the reading factor to the reading items were .75 to .82), and yet that the highly correlated nature of these two skills supported the reporting of a single total score (the coefficient between the two skill factors was .87). Since the reporting of two-skill scores and a total score was consistent with the way scores are reported in the revised TOEIC test, the study empirically supported the reporting practice of the revised TOEIC test.

According to the specifications of the TEAP test, a single higher-order or hierarchical factor is hypothesized to underlie performance on all four sections of the test. This is reflected in the test's construct definition of academic English proficiency and its breakdown of the test sections into four skills; the latter is also evidenced by the separate scores for the four-skill sections in the test's score report. Alternatively, TEAP examinees' performance may be hypothesized to be explained by distinctive factors of receptive (i.e., reading and listening) and productive (i.e., writing and speaking) skills.

Structure of L2 language ability

The higher-order factor structure of the TEAP test, as hypothesized above, is generally consistent with the literature on the structure of L2 language ability. Research in this area dates back to Oller's (1983) unitary trait hypothesis. Based on the principal component analysis of placement test data consisting of composition, vocabulary, grammar, phonology, and dictation, he stated that language proficiency was unitary and undividable into different skills and that it could be measured in its entirety, using cloze and dictation tests. A series of subsequent studies using confirmatory factor analysis, however, rejected the unitary trait hypothesis. For example, Bachman and Palmer (1982) analyzed three types of tests designed to measure grammatical, pragmatic, and sociolinguistic abilities and found that L2 ability was best explained by a higher-order model, with a general ability presiding over some specific first-order abilities. The higher-order model was also shown to best represent L2 ability in Bachman and Palmer (1989), Llosa (2007), Sawaki (2007), Sawaki et al. (2009), and Shin (2005).

The higher-order structure of L2 ability, however, has not always received support. Bachman and Palmer (1981) analyzed speaking and reading test data and found that L2 ability was best described as consisting of non-hierarchical, multiple components that correlated with each other. A similar finding was obtained in Sang et al (1986).

In sum, we can hypothesize that the ability or factor structure of the TEAP test is (a) hierarchically structured (based on, e.g., Sawaki 2007; Shin 2005), (b) non-hierarchical and closely correlated (based on, e.g., Bachman and Palmer 1981; Sang et al. 1986), or (c) separable into receptive (i.e., reading and listening) and productive (i.e., writing and speaking) components.

Current study

To further accumulate validity evidence for the TEAP test, we investigate the factor structure of the TEAP test in relation to the TOEFL iBT test. To the best of our knowledge, neither the factor structure of the TEAP test nor its relationship with the TOEFL iBT test has been examined. We decided to use the TOEFL iBT test as a criterion against which the TEAP test was compared, because it is a well-established measure of the four skills of English proficiency in an academic setting, with firm validity evidence reported (e.g., Chapelle et al. 2008; Educational Testing Service 2015c). Based on the studies reviewed above, we are interested in examining whether the ability or factor structure of the TEAP test is (a) hierarchically structured, (b) non-hierarchical and closely correlated, or (c) separable into receptive (i.e., reading and listening) and productive (i.e., writing and speaking) components. We are also interested in correlating the factor structure of the TEAP test with that of the TOEFL iBT test to examine the extent to which the two tests are related to each other. A strong relationship would suggest that the TEAP test measures academic English skills very well, whereas a weak relationship would suggest otherwise. Findings would contribute to not only the refinement or revision of the TEAP test's interpretation and use, but also the literature on the factor structure of L2 language ability and the correspondence between factor structures and score reporting practices. Two research questions are investigated herein:

(1) What is the factor structure of the TEAP test?

(2) How is the factor structure of the TEAP test related to that of the TOEFL iBT test?

Method

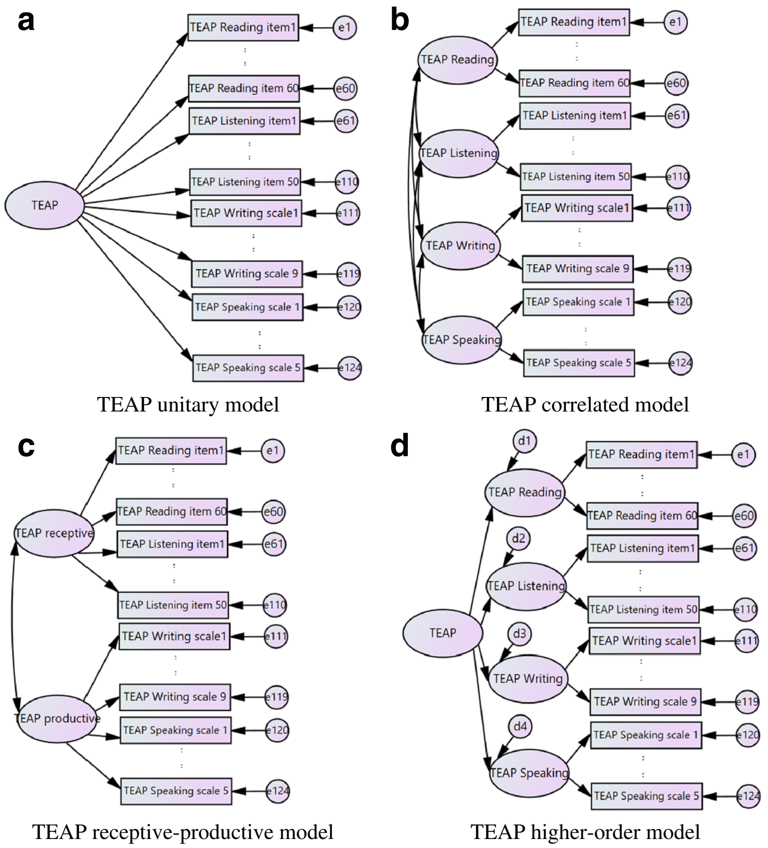
Data

Data were obtained from 100 first- and second-year undergraduate Japanese learners of English enrolled at a private university in Tokyo, taking both the TEAP and the TOEFL iBT tests. These students' English proficiency overall matched that of national or public university students with strong academic backgrounds. Stratified sampling was used to select participants according to proficiency level, to include learners with varying degrees of proficiency (Mean = 56.46, *SD* = 17.87, and score range = 18–95 in the TOEFL iBT; see descriptive statistics in Table 5). The TEAP test was administered in December 2013, as part of its validation study, under the supervision of its developer, the Eiken Foundation of Japan. One month before or after the administration of the TEAP test, the participants were required to take the TOEFL iBT test and report their section scores. The participants were paid upon completion of both tests. We had no missing data in the current study.

Analyses

The TEAP test's raw score at the item level and scaled score for each of the four skills were provided by the Eiken Foundation of Japan. The TOEFL iBT test data consisted of the scaled score for each of the four skills. The scaled scores—a maximum of 30 for each skill—were the only data available for analysis; the item-level TOEFL iBT data were not available for proprietary reasons.

The analyses were threefold, with the first and second analyses conducted for each test, and the third for the two test results combined. First, we examined the factor structure of the TEAP test (Research Question 1) by testing four models that hypothesized relationships among variables: a unitary model, a correlated model, a receptive-productive model, and a higher-order model. These models are presented in Fig. 1. In each figure, the rectangles represent the observed variables; the ovals, latent factors;



TEAP unitary model

TEAP correlated model

TEAP receptive-productive model

TEAP higher-order model

Fig. 1 Four item-level models tested for the TEAP test. All factors are correlated in Model 1B

and the circles, measurement errors or residuals. The data used for the observed variables are item-level dichotomous data for the Reading Section and Listening Section and the rating average of the two independent raters for four and five criteria for the Writing Section (four criteria for Task 1 [i.e., main ideas, coherence and cohesion, lexical range and accuracy, grammatical range and accuracy] and five criteria for Task 2 [i.e., main ideas, coherence, cohesion, lexical range and accuracy, and grammatical range and accuracy]) and for five criteria for the Speaking Section (pronunciation, lexical range and accuracy, grammatical range and accuracy, fluency, and interactional effectiveness). The Writing Section and Speaking Section data were rated on a 4-point scale, with 0 assigned to the lowest performance and 3 to the highest performance. When ratings diverged, a more experienced rater rated the performance of an examinee in question and finalized the rating (Nakatsuhara et al. 2014). According to the specifications of the TEAP test, the most preferable model was a higher-order model, but this had not been previously examined.

Second, to examine the factor structure of the TOEFL iBT test, we investigated whether a unitary model fit the data (see Fig. 2). It was not possible to test a higher-order factor model—the empirically supported factor structure of the TOEFL iBT test (see Fig. 7 in Sawaki et al. 2008, and Fig. 5 in Sawaki et al. 2009)—as we did not have access to the item-level data. Due to the small number of observed variables, resulting from the inclusion of only four, each of which represented one of the four skills, a higher-order factor model could not be constructed. However, a unitary factor model and a higher-order factor model are similar in that the four skills of reading, listening, writing, and speaking are subsumed in an underlying, higher-order ability. Otherwise stated, learners' four-skill performance is explained by a single, general ability that is hypothesized to be behind these four skills. Thus, although a higher-order model was not tested in the current study, it would be reasonable to argue that a favorable result for a unitary factor model, if obtained, indirectly supports a higher-order structure of the TOEFL iBT test.

Third, we combined and compared the best model from the TEAP test and the unitary model from the TOEFL iBT test to examine the extent to which the TEAP and the TOEFL iBT tests measured the same construct (Research Question 2; see Fig. 3).

Confirmatory factor analysis was used with the WLSMV estimator for Models 1A, 1B, 1C, 1D, and 3, all of which included the TEAP dichotomous data, and the maximum likelihood estimator for Model 2 in Mplus version 7.2 (Muthén and Muthén 1998–2014), to estimate model parameters. One of the factor loadings from each factor was fixed to 1, for scale identification. Model fit was evaluated by a non-significant chi-

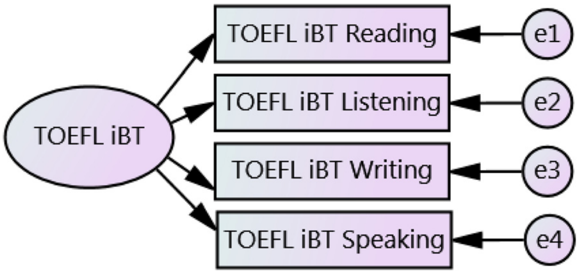


Fig. 2 Model 2: TOEFL iBT unitary model

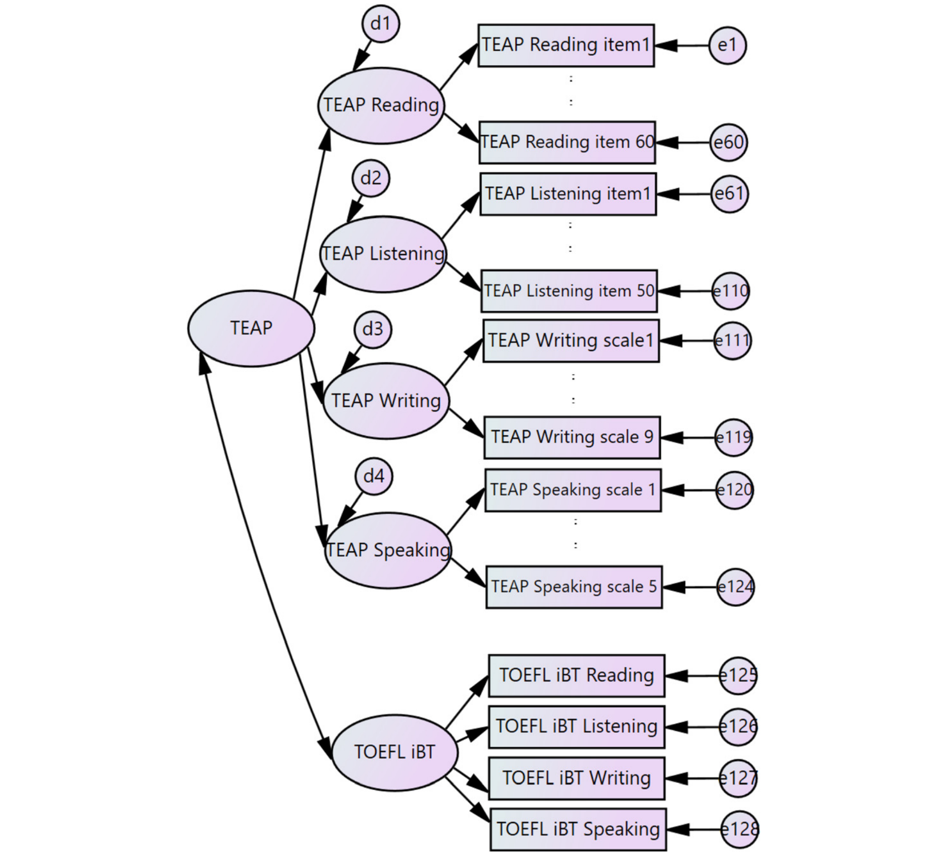


Fig. 3 Model 3: TEAP higher-order and TOEFL iBT unitary model

square (χ^2); a comparative fit index (CFI) and a Tucker-Lewis index (TLI) of .90 or higher; and a weighted root mean square residual (WRMR) of 1.0 or lower (Yu 2002); and a standardized root mean square residual (SRMR) of .08 or lower. Root mean square error of approximation (RMSEA) values of 0.05 or lower and 0.08 or lower, respectively, are often used as indicators of a close-fitting or reasonably-fitting model, based on Browne and Cudeck (1993). In the current study, the RMSEA of each model was reported, but not interpreted, because RMSEA tends to be too large in models with small degrees of freedom (*df*); this is particularly true with models with small sample sizes (Kenny et al. in press). For example, for $df = 2$, $N = 100$, and a cutoff of 0.05, as in the unitary model for the TEAP and the TOEFL iBT tests that will be reported in Table 6, the RMSEA is still larger than .05 at 28.7% of the time (see Table 1 of their article). For $df = 15$, $N = 100$, and a cutoff of 0.05, as in the combined unitary model for the TEAP and the TOEFL iBT tests, the RMSEA is still larger than .05 at 20.3%–25.5% of the time. Thus, we reported the RMSEA, but did not interpret it. Instead, we evaluated models based on the other fit indices. Chi-square difference tests were conducted to compare the four models for the TEAP test using the DIFFTEST option in Mplus, which is used with the WLSMV estimator.

To ensure that the current sample size of 100 was enough for all models to obtain adequate power and precision of parameter estimates, we conducted Monte Carlo studies following Muthén and Muthén (2002; see In'nami and Koizumi 2013, for concrete procedures). The results showed that the sample size was sufficient, except for the TEAP higher-order and TOEFL iBT unitary model (Model 3), which concerns Research Question 2. The covariance matrix of this model was not positive definite as will be reported below, which prevented us from calculating the power and precision of the parameter estimates.

As further reported on below, we also conducted item-parceling factor analysis as a follow-up by summing the item-level responses for each skill. For example, all 60 items in the Reading Section were summed. Models (see Fig. 4) were estimated using the maximum likelihood estimator. Additionally, a chi-square difference test was planned to compare models. This test was, however, not conducted, as elaborated on below. Please note that our main focus was on the item-level factor analysis, since it allowed us to examine if the data were best explained by the higher-order model. Since the latent variable covariance matrix was not positive definite and this might weaken support for the higher-order model as the best model to explain the data, we conducted item-parceling factor analysis to strengthen our findings.

Results

Item-level factor structure of the TEAP test

Table 2 shows fit statistics for the four hypothesized models of the TEAP test. The higher-order model (Model 1D) showed an overall good model fit with the data (CFI = .932, TLI = .931, RMSEA = .014 [.000, .022], and WRMR = .931). Good model-data fit was also obtained for the other three models.

The fit of the higher-order model was compared to the fit of the other three models using chi-square difference tests. Table 3 shows that the fit of the higher-order model was better than that of the other three models.

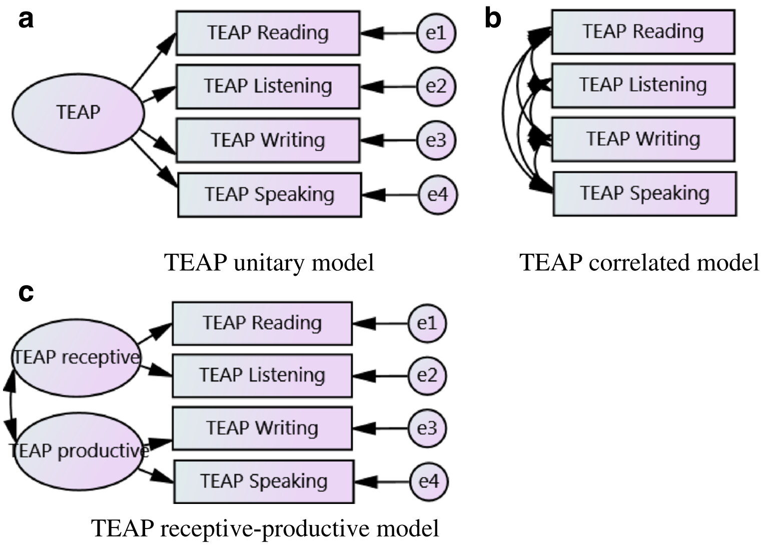


Fig. 4 Three item-parceling models tested for the TEAP test

Table 2 Fit Indices for the Models for Item-level Data

Model	χ^2	*df*	CFI	TLI	RMSEA [CI]	WRMR	SRMR
1A: TEAP unitary	7716.676*	7502	.907	.905	.017 [.004, .024]	.961	–
1B: TEAP correlated	7644.717	7496	.935	.934	.014 [.000, .022]	.927	–
1C: TEAP receptive-productive	7686.215	7501	.920	.918	.016 [.000, .023]	.946	–
1D: TEAP higher-order	7653.733	7498	.932	.931	.014 [.000, .022]	.931	–
2: TOEFL iBT unitary	15.130*	2	.944	.833	.256 [.146, .384]	–	.039
3: TEAP higher-order and TOEFL iBT unitary	8158.233	7995	.938	.937	.014 [.000, .022]	.917	–

Note. df degrees of freedom, *CI* 90% confidence interval. *$p < .05$

Factor structure of the TOEFL iBT test

Table 2 shows fit statistics for the unitary model (Model 2) of the TOEFL iBT test. Although the chi-square statistic was statistically significant ($\chi^2 = 15.130$, $df = 2$, $p < .001$), other statistics showed an overall good model fit with the data (CFI = .944 and SRMR = .039). Nevertheless, its slightly lower TLI (.833) suggests that the fit was somewhat compromised and that the model warrants less confidence. The RMSEA was high (.256 [.146, .384]), but not interpreted as explained in the Analyses section.

Comparison of the TEAP test with the TOEFL iBT test using item-level data

Table 2 shows fit statistics for the TEAP higher-order model and the TOEFL iBT unitary model, as combined and compared (Model 3). This model showed an overall good model fit with the data (CFI = .938, TLI = .937, RMSEA = .014 [.000, .022], and WRMR = .917).

The model is presented in Table 4. Since the parameter estimates were all standardized, we can directly compare them in order to examine their relative strength as indicators of the relationships between the factors and the observed variables, as well as between the factors themselves. The path coefficients of the TEAP and TOEFL iBT factors to each of the observed variables (Reading, Listening, Writing, and Speaking) were on average medium to high (mean = .514 [range = .122 to .960] for the TEAP Reading, mean = .517 [range = .093 to .880] for the TEAP Listening, mean = .747 [range = .426 to .901] for the TEAP Writing, mean = .835 [range = .724 to .938] for the TEAP Speaking; and mean = .797 [range = .718 to .872] for the TOEFL iBT). This suggests that both tests measured the four skills of reading, listening, writing, and speaking very well. Overall, the TEAP higher-order factor was highly loaded on the TEAP four-skill factors, ranging from .693 to .925, suggesting the hierarchical structure of the test. The coefficient between the TEAP higher-order factor and the TOEFL iBT factor was high (1.005), exceeding .90, which was considered to indicate that the two factors were not

Table 3 Chi-Square Difference Test Results for the TEAP Higher-order Model Versus the Three Alternative Models

	χ^2 difference	*df* difference
Vs. Unitary	87.055*	4
Vs. Correlated	10.172*	2
Vs. Receptive-productive	43.375*	3

Note. df degrees of freedom. *$p < .05$

Table 4 Standardized Parameter Estimates for the TEAP Higher-order Model with the TOEFL iBT Unitary Model (Model 3; Fig. 3)

	TEAP				TOEFL				TEAP higher-order	Error	R^2
Item	R	L	W	S	R	L	W	S			
RITEM1	.696									.119	.485
RITEM2	.442									.121	.195
RITEM3	.247									.139	.061
RITEM4	.795									.062	.631
RITEM5	.582									.100	.339
RITEM6	.246									.149	.060
RITEM7	.651									.119	.424
RITEM8	.778									.086	.605
RITEM9	.343									.122	.118
RITEM10	.618									.102	.382
RITEM11	.245									.132	.060
RITEM12	.836									.076	.700
RITEM13	.406									.116	.165
RITEM14	.724									.079	.525
RITEM15	.789									.106	.623
RITEM16	.713									.081	.509
RITEM17	.485									.113	.236
RITEM18	.960									.052	.921
RITEM19	.445									.134	.198
RITEM20	.759									.083	.576
RITEM21	.242									.124	.058
RITEM22	.597									.121	.356
RITEM23	.122									.162	.015
RITEM24	.669									.102	.447
RITEM25	.273									.125	.075
RITEM26	.527									.114	.277
RITEM27	.460									.128	.211
RITEM28	.580									.113	.336
RITEM29	.159									.136	.025
RITEM30	.254									.154	.064
RITEM31	.608									.134	.369
RITEM32	.586									.114	.344
RITEM33	.413									.128	.170
RITEM34	.812									.068	.659
RITEM35	.511									.115	.261
RITEM36	.282									.129	.079
RITEM37	.644									.098	.414
RITEM38	.500									.116	.250
RITEM39	.650									.098	.422
RITEM40	.391									.113	.153
RITEM41	.463									.107	.214

Table 4 Standardized Parameter Estimates for the TEAP Higher-order Model with the TOEFL iBT Unitary Model (Model 3; Fig. 3) *(Continued)*

RITEM42	.331		.146	.110
RITEM43	.525		.113	.275
RITEM44	.120		.131	.014
RITEM45	.436		.142	.190
RITEM46	.605		.103	.365
RITEM47	.671		.087	.450
RITEM48	.495		.106	.245
RITEM49	.354		.121	.126
RITEM50	.346		.116	.120
RITEM51	.427		.115	.183
RITEM52	.649		.094	.421
RITEM53	.430		.135	.185
RITEM54	.682		.086	.465
RITEM55	.612		.093	.375
RITEM56	.639		.088	.409
RITEM57	.375		.122	.140
RITEM58	.673		.098	.453
RITEM59	.706		.088	.498
RITEM60	.519		.112	.269
LITEM1		.812	.099	.659
LITEM2		.787	.106	.620
LITEM3		.806	.126	.650
LITEM4		.430	.123	.185
LITEM5		.587	.119	.344
LITEM6		.596	.094	.355
LITEM7		.771	.070	.594
LITEM8		.438	.113	.192
LITEM9		.674	.088	.454
LITEM10		.277	.147	.076
LITEM11		.702	.093	.493
LITEM12		.598	.101	.358
LITEM13		.692	.085	.479
LITEM14		.557	.109	.311
LITEM15		.752	.119	.565
LITEM16		.560	.113	.314
LITEM17		.805	.071	.648
LITEM18		.306	.130	.094
LITEM19		.235	.131	.055
LITEM20		.452	.115	.205
LITEM21		.880	.060	.774
LITEM22		.324	.122	.105
LITEM23		.335	.116	.112
LITEM24		.296	.121	.088

Table 4 Standardized Parameter Estimates for the TEAP Higher-order Model with the TOEFL iBT Unitary Model (Model 3; Fig. 3) *(Continued)*

LITEM25	.419						.107	.175
LITEM26	.469						.132	.220
LITEM27	.374						.128	.140
LITEM28	.158						.131	.025
LITEM29	.538						.127	.290
LITEM30	.427						.117	.182
LITEM31	.698						.098	.487
LITEM32	.661						.085	.437
LITEM33	.385						.122	.149
LITEM34	.470						.108	.221
LITEM35	.230						.127	.053
LITEM36	.532						.099	.283
LITEM37	.524						.103	.274
LITEM38	.357						.125	.128
LITEM39	.280						.138	.079
LITEM40	.402						.118	.162
LITEM41	.457						.109	.209
LITEM42	.307						.121	.094
LITEM43	.623						.094	.388
LITEM44	.780						.069	.609
LITEM45	.834						.077	.695
LITEM46	.715						.095	.511
LITEM47	.504						.107	.254
LITEM48	.445						.111	.198
LITEM49	.505						.103	.255
LITEM50	.093						.133	.009
WITEM1		.426					.076	.182
WITEM2		.790					.053	.625
WITEM3		.745					.060	.555
WITEM4		.772					.050	.596
WITEM5		.737					.055	.544
WITEM6		.735					.062	.540
WITEM7		.901					.041	.812
WITEM8		.832					.039	.692
WITEM9		.787					.053	.619
SITEM1			.724				.076	.525
SITEM2			.938				.048	.880
SITEM3			.911				.047	.831
SITEM4			.794				.062	.631
SITEM5			.810				.066	.656
TOEFL_R				.718			.053	.516
TOEFL_L					.872		.034	.760
TOEFL_W						.856	.033	.733

Table 4 Standardized Parameter Estimates for the TEAP Higher-order Model with the TOEFL iBT Unitary Model (Model 3; Fig. 3) *(Continued)*

TOEFL_S	.742		.050	.550
Higher-order factor loadings				
TEAP_R		.901	.028	.812
TEAP_L		.925	.033	.855
TEAP_W		.722	.052	.521
TEAP_S		.693	.062	.480
Interfactor coefficient				
TOEFL iBT		1.005	.021	

Note. Each item is labeled by skill and item number. For example, RITEM1, LITEM1, WITEM1, and SITEM1 refer to Item 1 in the Reading, Listening, Writing, or Speaking Sections, respectively. TOEFL_R, TOEFL_L, TOEFL_W, and TOEFL_S refer to the sum score in the Reading, Listening, Writing, and Speaking Sections, respectively. The paths from the RITEM1, LITEM1, WITEM1, SITEM1, and TOEFL_R are fixed to 1 for identification. Factor loadings for the RITEM3, RITEM11, RITEM21, RITEM23, RITEM29, RITEM30, RITEM44, LITEM10, LITEM19, LITEM28, LITEM35, and LITEM50 are statistically nonsignificant. All other factor loadings are statistically significant

distinct from each other (Sawaki et al. 2009). This suggests that the TEAP and TOEFL iBT tests measured a very similar construct.

Nevertheless, this support for the TEAP test must be considered with the Mplus output warning that the latent variable covariance matrix was not positive definite. This could be due to: (a) a negative variance/residual variance for a latent variable, (b) a correlation greater or equal to 1 between two latent variables, or (c) a linear dependency among more than two latent variables. Of these, (a) was unlikely as we had no such variances, while (b) and (c) were likely since the estimated relationship between the TEAP higher-order factor and the TOEFL iBT factor was 1.005. However, this is a regression coefficient and a coefficient of more than 1 could occur. According to Jöreskog (1999; also see Deegan 1978), "if the factors are correlated (oblique), the factor loadings are regression coefficients and not correlations and as such they can be larger than one in magnitude" (p. 1). Thus, the coefficient of 1.005 might have made the matrix not positive definite, but might nevertheless not invalidate our findings.

However, having the not positive definite matrix was worrisome, for which reason we decided to conduct further analysis to replicate the relationship between the TEAP and TOEFL iBT tests. We conducted item-parceling factor analysis by summing item-level responses for each skill. If the close relationship between the TEAP and TOEFL iBT tests was again observed, that would render the plausible threat of the not positive definite matrix harmless and strengthen our findings.

Item-parceling factor structure of the TEAP test

Descriptive statistics in Table 5 show that all the values for skewness and kurtosis were within |3.30| (z score at $p < .01$), suggesting univariate normality of the data (e.g., Tabachnick and Fidell 2007). Mardia's normalized estimates for skewness and kurtosis were both statistically non-significant ($p > .05$), suggesting multivariate normality of the data. The correlation matrix is provided in the Appendix.

Table 6 shows fit statistics for the three hypothesized models of the TEAP test. For the unitary model (Model 4A), although the chi-square statistic was statistically significant ($\chi^2 = 11.101$, $df = 2$, $p < .001$), other statistics showed an overall good model fit with

Table 5 Descriptive Statistics for Item-parceling Data

	Mean (%)	*SD*	Minimum	Maximum	Skewness	Kurtosis	Possible maximum score
TEAP Reading	39.170 (65.283)	10.215	10	57	−0.538	−0.346	60 (100[a])
TEAP Listening	31.670 (63.340)	8.607	9	48	−0.258	−0.579	50 (100[a])
TEAP Writing	16.550 (61.296)	5.062	4	27	−0.549	0.272	27 (100[a])
TEAP Speaking	9.010 (60.067)	2.973	1	15	0.270	−0.291	15 (100[a])
TEAP Total	96.400 (63.421)	23.110	3	145	−0.290	−0.543	152 (400[a])
TOEFL iBT Reading	14.150 (47.167)	5.387	1	25	−0.307	−0.403	30[a]
TOEFL iBT Listening	13.130 (43.767)	5.729	1	25	0.224	−0.422	30[a]
TOEFL iBT Writing	15.090 (50.300)	4.797	5	25	−0.001	−0.570	30[a]
TOEFL iBT Speaking	14.090 (46.967)	4.987	3	24	0.087	−0.390	30[a]
TOEFL iBT Total	56.460 (47.050)	17.870	18	95	0.033	−0.628	120[a]

Note. [a]Scaled score. We analyzed raw scores of the TEAP test and scaled scores of the TOEFL iBT test

the data (CFI = .951 and SRMR = .033). Nevertheless, its slightly lower TLI (.854) suggests that the fit was somewhat compromised and that the model warrants less confidence. This could have been due to the average low correlation (*r* = .689) among the observed variables (see the correlation matrix in Appendix), since TLI will be low if the average correlation in the data is low (Kenny 2014). In Appendix, note the particularly low correlations between TEAP Reading and TEAP Speaking (*r* = .497), TEAP Writing and TOEFL iBT Speaking (*r* = .491), TEAP Speaking and TOEFL iBT Reading (*r* = .335), and TOEFL iBT Reading and TOEFL iBT Speaking (*r* = .419). The cause of these low correlations was not clear, except the suggestion that they measured somewhat different constructs. Given this and the fact that the TLI was only slightly lower, we considered the unitary model still useful. Finally, the RMSEA was high (.213 [90% confidence interval: .103, .343]), but not interpreted because, as we discussed earlier, it tends to be inflated in models with small degrees of freedom and with small sample sizes (Kenny et al., in press). Collectively, the unitary model fit the data well, overall.

The fit of the correlated model (Model 4B) could not be examined, since it had a degree of freedom of zero and all fit indices were perfect, accordingly (i.e., the model was just-identified). This neither indicates the extent to which the correlated model was useful in explaining the current data, nor the adequacy of comparing the correlated model with the other models. This led us to exclude the model from subsequent analyses. The receptive-productive model (Model 4C) produced a statistically significant

Table 6 Fit Indices for the Models for Item-parceling Data

Model	χ^2	*df*	CFI	TLI	RMSEA [CI]	WRMR	SRMR
4A: TEAP unitary	11.101*	2	.951	.854	.213 [.103, .343]	–	.033
4B: TEAP correlated	0.000	0	1.000	1.000	.000 [.000, .000]	–	.000
4C: TEAP receptive-productive	9.766*	1	.953	.718	.296 [.148, .477]	–	.030
5: TEAP unitary and TOEFL iBT unitary	37.692*	15	.963	.930	.123 [.074, .173]	–	.040

Note. df = degrees of freedom. CI = 90% confidence interval. *$p < .05$. Before constructing Models 4A, 4B, and 4C, we conducted item-level factor analysis to examine whether each skill was unidimensional, to judge the adequacy of aggregating the item-level score to form a total section score for each skill. For example, we conducted item-level factor analysis on 60 items for the Reading Section, to examine if they were unidimensional, and thereby warrant being combined into a single score. The results supported the unidimensionality of each skill (i.e., for reading, χ^2 = 1803.004, *df* = 1710, *p* = .0578, CFI = .930, TLI = .928, RMSEA [CI] = .023 [.000, .034], and WRMR = .941) chi-square statistic ($\chi^2 = 9.766$, $df = 1$, $p < .001$), and yet an overall good model fit with the data (CFI = .953 and SRMR = .030). It produced, however, a considerably low TLI (.718), and this was a cause of concern. Although this could have been again due to the low correlations among the observed variables, as TLI is a function of the average correlations in the data, the value was so low that we considered the receptive-productive model unhelpful in explaining the data. In the end, the correlated model was excluded and the receptive-productive model was rejected.

In sum, although we planned to conduct a chi-square difference test to compare the unitary model with the correlated model and the receptive-productive model, we no longer needed to do so, since the latter two models were found to be unsatisfactory. The unitary model best represented the factor structure of the TEAP test.

Comparison of the TEAP test with the TOEFL iBT test using item-parceling data

Table 6 shows fit statistics for the two unitary models (Model 5) of both tests combined and compared. Although the chi-square statistic was statistically significant ($\chi^2 = 37.692$, $df = 15$, $p < .001$), all other statistics showed an overall good model fit with the data (CFI = .963, TLI = .930, and SRMR = .040), except RMSEA (.123 [.074, .173]).

The TEAP unitary and TOEFL iBT unitary model (Model 5) is shown in Fig. 5. Note that the measurement error variances of the same skill were allowed to covary (e.g., TEAP Reading and TOEFL iBT Reading). This was because both measures were designed to test the four skills of reading, listening, writing, and speaking in an academic setting and included constructs, task formats, or both, that were similar across the measures. Thus, it was sensible to hypothesize that measurement errors for each skill were related to each other.

The parameter estimates in Fig. 5 were all standardized, and we can directly compare them. The path coefficients of the TEAP and TOEFL iBT factors to the observed variables (Reading, Listening, Writing, and Speaking each) to the corresponding TEAP and

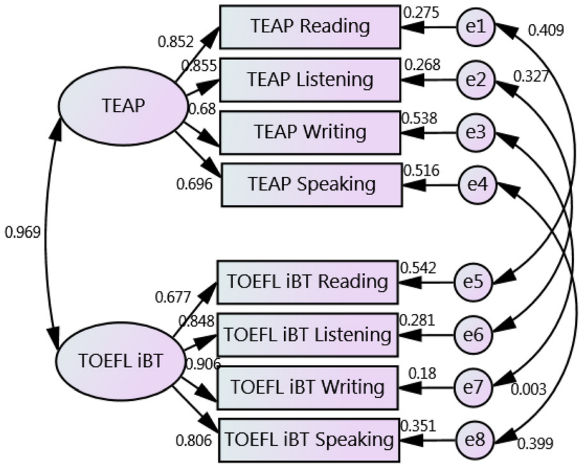


Fig. 5 Model 5: TEAP unitary and TOEFL iBT unitary model. All values are standardized. The paths from the TEAP factor to the TEAP Reading and from the TOEFL iBT factor to the TOEFL iBT Reading are fixed to 1 for identification. All other factor loadings are statistically significant. R^2 is .725, .732, .462, and .484 for the TEAP Reading, Listening, Writing, and Speaking, respectively; .458, .719, .820, and .649 for the TOEFL Reading, Listening, Writing, and Speaking, respectively

TOEFL iBT factors were medium to high (from .677 to .906). This suggests that both tests measured the four skills of reading, listening, writing, and speaking very well. The coefficient between the two factors was also high (.969), which exceeded .90 and was considered to indicate that the two factors were not distinct from each other (Sawaki et al. 2009). This suggests that the TEAP and TOEFL iBT tests measured a very similar construct. The results for the item-parceling factor analysis supported those for the item-level factor analysis.

The error correlations of the same skills between the TEAP and the TOEFL iBT tests were generally low ($r = .327$ to .409), except for writing, to a negligible degree ($r = .003$). The low error correlation could be explained by differences in the abilities tested and difficulty levels across the tests. The TEAP Writing Section consists of two tasks, with (a) one requiring examinees to summarize a reading text and (b) the other to summarize main points and write their opinions, based on multiple reading texts. In contrast, the TOEFL iBT Writing section requires examinees (c) to summarize reading and *listening* texts and (d) to write their opinions about the topic assigned, *without* reading texts. Thus, the TEAP Writing and the TOEFL iBT Writing seem to measure slightly different abilities, and this might have led to the low correlation of the measurement errors for writing between the two tests. The other three-skill sections had clear similarities—the same test formats in reading and listening (i.e., all multiple-choice type questions) and similar abilities tested in speaking (i.e., describing familiar topics, stating opinions, and speaking based on listening input).

Discussion and conclusion

In order to further accumulate validity evidence for the TEAP test—a recently developed English test for Japanese university applicants, designed to measure four skills of reading, listening, writing, and speaking in an academic context—we examined the factor structure of the TEAP test in relation to that of the TOEFL iBT test—a more established instrument of the four skills in an academic setting—to examine the extent to which the TEAP test measures a construct similar to the one measured by the TOEFL iBT test. Confirmatory factor analysis was used on data collected from 100 Japanese EFL students taking both tests. Research Question 1 asked what the factor structure of the TEAP test was. More specifically, we examined whether the unitary, correlated, receptive-productive, or higher-order model assumed to underlie performance on the TEAP test fit the data best. Of these four item-level response models (Models 1A to 1D), the higher-order model explained the TEAP test data best as shown in Table 3.

As a follow-up analysis, three item-parceling models (Models 4A to 4C) were tested. As shown in Table 6, the unitary model was selected as the best model for the TEAP test, although it produced a slightly lower TLI value. The correlated model was excluded due to its inability to be tested using fit indices, since its degree of freedom was zero. The receptive-productive model was rejected due to its considerably low TLI value, suggesting that this model does not reflect the construct measured in the TEAP test well. Selecting the item-parceling unitary model as the best-fitting model (Models 4A) was consistent with the aforementioned item-level higher-order model (Models 1D). This indicates that the TEAP test measures the four skills of reading, listening, writing, and speaking well and that these skills could be conceptualized as reflecting a single academic proficiency. This supports the appropriateness of the constructs, as defined and operationalized in the TEAP test specification.

Research Question 2 asked how the factor structure of the TEAP test was related to that of the TOEFL iBT test. As reported in Table 2, this model (Model 3) showed an overall good model fit with the data. The path coefficients of the observed variables to the corresponding TEAP and TOEFL iBT factors were on average medium to high, suggesting that the four skills in each test were measured well. Overall, the TEAP higher-order factor was highly loaded on the TEAP four-skill factors, suggesting the hierarchical ability structure of the test. Most importantly, the high coefficient between the TEAP higher-order factor and the TOEFL iBT factor (1.005) exceeded .90, indicating the inseparability of the two factors (Sawaki et al. 2009) and the measurement of a very similar construct in these two tests. As the TOEFL iBT test has been a well-established measure of the four skills of English proficiency in the academic setting, the current close relationship between the TEAP and TOEFL iBT tests further suggests that the TEAP test measures the intended construct of the four skills in academic settings very well.

A follow-up item-parceling analysis provided the same result. As seen in Table 6, the two unitary models of both tests combined were found to fit the data well. The final model (Model 5) had medium to high path coefficients, from the factors to the observed variables (from .677 to .906), suggesting that the four skills in each test were measured well. The close relationship between the two factors (.969) suggests that they are separate, yet related closely enough to be combined into one factor (Sawaki et al. 2009) and, more importantly, that the TEAP and TOEFL iBT tests measure a very similar construct. This provides positive evidence of the concurrent validity of the TEAP test, since, despite its short history, its strong relationship with the TOEFL iBT, which has been supported by positive validity evidence through extensive research, seems to support the use of the TEAP test as an indicator of English skills in an academic context. We hasten to add, however, that the close relationship between the two tests does not necessarily suggest the replaceability of one with another. As shown in Table 5's depiction of the percentage of correct answers for each section and the total score of the tests, the percentage of correct answers on the TEAP test is higher than that on the TOEFL iBT test, suggesting that the TEAP test is easier than the TOEFL iBT test, as intended. Further, the TEAP test differs in content from the TOEFL iBT test; the former is designed for Japanese high school students who have learned English based on curriculum guidelines. Thus, the TEAP test is closely related to the TOEFL iBT test, but differs in terms of difficulty and content.

Implications and future research

The findings regarding the factor structure of the TEAP test and its relationship with the factor structure of the TOEFL iBT test have three main implications. First, the presence of distinctive observed variables for each skill in the TEAP test supports the reporting of separate scores for each skill. This reporting format is in accordance with that used in the TEAP test, and, thus, the current results provide empirical support for the reporting practice adopted by the TEAP test. Moreover, the satisfactory model-data fit of the unitary factor model suggested the distinct, but relatively highly related nature of these four skills and support such single-score reporting.

Second, the close relationship between the TEAP and the TOEFL iBT tests (1.005 and .969 for the item-level and item-parceling analyses, respectively) suggests that the TEAP and the TOEFL iBT tests measure a very similar construct. This does not mean that the TEAP test is not necessary or that it can be replaced with the TOEFL iBT test. The TEAP test was intended for Japanese university applicants, with its difficulty level and content designed to be appropriate for them, in accordance with Japanese high school curriculum guidelines.

Third, the current study contributes to discussions on the factor structure of L2 language ability. Our study found that the item-level higher-order structure of the TEAP test concurred with previous studies on the higher-order factor structure of L2 language ability (Bachman and Palmer 1989; Llosa 2007; Sawaki 2007; Sawaki et al. 2009; and Shin 2005). The item-parceling unitary structure indirectly supported such higher-order structure.

Further research is needed in three areas. First, the unavailability of the item-level data precluded modeling a higher-order model for the TOEFL iBT test. With access to such data and with the replication of the current study with a larger sample size, we can gain stronger evidence of the relationship between the TEAP and the TOEFL iBT tests. Second, we used the TEAP data collected from undergraduate students at a private university in Japan. While our proficiency-stratified sampling was reasonably successful in recruiting learners of a wide range of proficiency (score range = 18–95 in the TOEFL iBT; see descriptive statistics in Table 5), we were not able to include advanced learners who had obtained TOEFL iBT scores of more than 95. The lack of advanced learners in the study sample might have affected the relationship between the TEAP and TOEFL iBT scores. Also note that the mean TOEFL iBT score of 56.46 for the current examinees was higher than the expected TOEFL iBT score of 24 to 30 from Japanese high school students based on MEXT (2014b). The current study finding must be replicated with more diverse target populations of test-takers. Third, our data were drawn from one of the several existing forms of both tests. Although all forms are designed to be equivalent, in terms of their content and difficulty, it remains to be seen as to whether the factor structure of the TEAP test and the close relationship between the TEAP and the TOEFL iBT tests will be supported in other forms of the two tests.

Table 7 Pearson's Product–moment Correlation Matrix (*N* = 100)

	1	2	3	4	5	6	7	8	9	10
1: TEAP Reading	–									
2: TEAP Listening	.746	–								
3: TEAP Writing	.614	.549	–							
4: TEAP Speaking	.497	.646	.517	–						
5: TEAP Total	.918	.905	.761	.702	–					
6: TOEFL iBT Reading	.722	.572	.534	.335	.692	–				
7: TOEFL iBT Listening	.765	.796	.554	.557	.828	.602	–			
8: TOEFL iBT Writing	.748	.727	.594	.638	.814	.642	.736	–		
9: TOEFL iBT Speaking	.596	.683	.491	.706	.716	.419	.690	.771	–	
10: TOEFL iBT Total	.830	.813	.635	.648	.892	.784	.892	.913	.833	–

Note. All correlations are significant at the .01 level (2-tailed)

Appendix

Competing interests
KN is employed by the Eiken Foundation of Japan, which develops and administers the TEAP test.

Authors' contributions
YI performed the statistical analysis and drafted the manuscript. RK assisted YI in performing the statistical analysis and drafted the manuscript. KN participated in the design of the study, collected the data, and assisted YI in drafting the manuscript. All authors read and approved the final manuscript.

Author details
[1]Chuo University, Tokyo, Japan. [2]Juntendo University, Chiba, Japan. [3]Eiken Foundation of Japan, Tokyo, Japan.

Received: 31 October 2015 Accepted: 10 February 2016
Published online: 25 February 2016

References

ALC Press. (2015). *The Standard Speaking Test*. Retrieved from http://tsst.alc.co.jp/sst/e/index.html.

Bachman, L. F., & Palmer, A. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 67–86.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449–465.

Bachman, L. F., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*, 14–29.

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English two-way Immersion program. *Language Testing, 15*, 380–414.

Benesse. (2000–2015a). *GTEC CBT*. Retrieved from http://www.benesse-gtec.com/cbt/en

Benesse. (2000–2015b). *GTEC for STUDENTS*. Retrieved from http://www.benesse-gtec.com/fs/

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons.

British Council, IDP, IELTS Australia, & Cambridge English Language Assessment. (n.d.). *IELTS*. Retrieved from http://www.ielts.org/

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills: Sage.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*™. New York: Routledge.

Council of Europe. (2014). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Deegan, J., Jr. (1978). On the occurrence of standardized regression coefficients greater than one. *Educational and Psychological Measurement, 38*, 873–888.

Educational Testing Service. (2015a). *About the TOEFL iBT® Test*. Retrieve from https://www.ets.org/toefl/ibt/about.

Educational Testing Service. (2015b). *About the TOEFL Junior® Test*. Retrieve from https://www.ets.org/toefl_junior/about.

Educational Testing Service. (2015c). *TOEFL® research reports*. Retrieved from https://www.ets.org/toefl/research/archives/research_report/.

Educational Testing Service. (2015d). *The TOEIC® Test*. Retrieved from https://www.ets.org/toeic.

Eiken Foundation of Japan. (n.d.a). EIKEN Tests. Retrieved from http://www.eiken.or.jp/eiken/en/eiken-tests/

Eiken Foundation of Japan. (n.d.b.). *Mondai kousei* [Sections of the TEAP]. Retrieved from http://www.eiken.or.jp/teap/construct/

Eiken Foundation of Japan. (n.d.c). *TEAP kenkyu report* [TEAP research reports]. Retrieved from http://www.eiken.or.jp/teap/group/report.html

Green, A. (2014). *The Test of English for Academic Purposes (TEAP) impact study: Report 1—Preliminary questionnaires to Japanese high school students and teachers*. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*, 131–152.

In'nami, Y., & Koizumi, R. (2013). Review of sample size for structural equation models in second language testing and learning research: A Monte Carlo approach. *International Journal of Testing, 13*, 329–353.

Japan Institute for Educational Measurement. (n.d.). *Computerized Assessment System for English Communication (CASEC)*. Retrieved from http://global.casec.com/

Jöreskog, K. G. (1999). *How large can a standardized coefficient be?* Retrieved from http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf.

Kenny, D. A. (2014). *Measuring model fit*. Retrieved from http://davidakenny.net/cm/fit.htm.

Kenny, D. A, Kaniskan, B, McCoach, D. B. (in press). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*, 489–515.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.

Ministry of Education, Culture, Sports, Science & Technology (MEXT). (2014a). *On integrated reforms in high school and university education and university entrance examination aimed at realizing a high school and university articulation system appropriate for a new era (Report)*. Retrieved from http://www.mext.go.jp/english/topics/1356088.htm.

Ministry of Education, Culture, Sports, Science & Technology (MEXT). (2014b). *Heisei 26 nendo eigokyoiku kaizen no tameno eigoryoku chosa jigy hokoku [The English proficiency of Japanese high school students]*. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1358258.htm.

Muthén, L. K., & Muthén, B. O. (1998–2014). *Mplus [Computer software]*. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599–620.

Nakamura, K (2014). *Examination of possible consequences of a new test within the context of university entrance exam reform in Japan*. Paper presented at the 36th Language Testing Research Colloquium, VU University Amsterdam, the Netherland. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_ltrcpresentation20140620.pdf

Nakatsuhara, F. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) speaking test for Japanese university entrants—Study 1 & Study 2*. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf.

Nakatsuhara, F., Joyce, D., & Fouts, T. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) speaking test for Japanese university entrants—Study 3 & Study 4*. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_speaking_report2.pdf.

National Center for University Entrance Examinations. (2015). *Shiganshasuu (kakutei) ni tsuite [The number of applicants]*. Retrieved from http://www.dnc.ac.jp/albums/abm.php?f=abm00004641.pdf&n=%E5%BF%97%E9%A1%98%E8%80%85%E6%95%B0%EF%BC%88%E7%A2%BA%E5%AE%9A%EF%BC%89%E3%81%AB%E3%81%A4%E3%81%84%E3%81%A6.pdf.

Ockey, G. J. (2009). The effects of group members' personalities on a test-taker's L2 group oral discussion test scores. *Language Testing, 26*, 161–186.

Oller, J. W., Jr. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller Jr. (Ed.), *Issues in language testing research* (pp. 3–10). Rowley: Newbury House.

Pearson. (2014). *PTE Academic*. Retrieved from http://pearsonpte.com/.

Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J., & Roeder, P. M. (1986). Models of second language competence: A structural equation approach. *Language Testing, 3*, 54–79.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*, 355–390.

Sawaki, Y, Stricker, L. J, Oranje, A. H. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample*. TOEFL iBT Research Report TOEFLiBT-04. Retrieved from https://www.ets.org/Media/Research/pdf/RR-08-09.pdf

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*, 5–30.

Shimomura, H. (2014). *Statement by Minister of Education, Culture, Sports, Science and Technology of Japan on the October 12 International New York Times article "Japan's Divided Education Strategy."*. Retrieved from http://www.mext.go.jp/english/topics/1353287.htm.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*, 31–57.

Sophia University. (n.d.a). *2015 nendo (heisei 27 nendo) Jouchidaigaku ippannyuugakushiken no henkou ni tsuite (yokoku)* [Changes in Sophia University's entrance examination system starting 2015]. Retrieved from http://www.sophia.ac.jp/jpn/admissions/gakubu_ad/gakubu_news/20130524/gakubunews20130524?kind=0

Sophia University. (n.d.b). *TEAP ni tsuite* [About TEAP]. Retrieved from http://www.sophia.ac.jp/jpn/admissions/gakubu_kanren/teap

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights: Allyn and Bacon.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology (ETS Research Rep. No. RR-08-34; TOEFL iBT Research Rep. No. TOEFLiBT-06)*. Princeton: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-08-34.pdf.

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese university entrants*. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_rlspecreview_report.pdf.

University of Cambridge Local Examinations Syndicate. (2015). *Cambridge English exams*. Retrieved from http://www.cambridgeenglish.org/exams/.

Weir, C. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) writing test for Japanese university entrants*. Retrieved from http://www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Doctoral dissertation)*. Retrieved from http://www.statmodel.com/download/Yudissertation.pdf.